# Exam for Machine Learning Python Lab

Consider the file provided with the assignment, explore the data, drop the columns that you consider useless for the clustering and find the best clustering scheme considering only the relevant columns.
The solution must be produced as a Python Notebook, assuming that the dataset is in the same folder as the notebook.
You can use only the computers of the lab, you cannot use any other device, you cannot use email or any other messaging tool. You can use only the websites accessible through the computers of the lab.
The notebook must operate as follows:

1. Load the data file and explore the data, showing size, data descriptions, data distributions with boxplot, pairplots ..... **1pt**

2. Comment the exploration of step 1 pointing out if there are imbalanced distributions, outliers, missing values, features that seem not to be relevant for clustering ..... **1pt**

3. Drop the columns that are not relevant for the clustering operation, if any, and explain why you do that.
   Deal with missing values, if any.
   Transform the fields with type "object" and only two distinct values into '0/1' with `OrdinalEncoder`
   Transform the other "object" fields with `OneHotEncoder`
   Transform the numeric fields with `MinMaxScaler` ..... **6pt**

4. find the best clustering scheme with KMeans, show the silhouette plots of clusters, show the distribution of the resulting cluster labels (e.g. histogram or pie plot) ..... **4pt**

5. find the best clustering scheme with Agglomerative Clustering or DBSCAN (your choice) show the silhouette plots of clusters, show the distribution of the resulting cluster labels (e.g. histogram or pie plot) **2pt**

6. Compare the similarities of the two schemes with the `adjusted_rand_score` and comment the results ..... **2pt**

*Quality of the code* ..... **4pt**

- Include appropriate comments with reference to the numbered requirements
- Useless cells, pieces of code and non-required output will be penalised
- Remove the code you use for testing and inspecting the variables during the development
- Naming style of variables must be uniform and in English
- Bad indentation and messy code will be penalised
- Non generalised solution, such as three sequential statements with the same kind of operation instead of a loop, will be penalised

Additional directions, the assignments not compliant with the rules below will not be considered:

- The notebook name must be `yourworkplace_youremailusername.ipynb` in lowercase letters
  E.G. if your worplace is `lab9_35` and your email is `mario.rossi45@studio.unibo.it`, the notebook filename will be `lab9_35_mario.rossi45.ipynb`
- The solution must directly access the data in the same folder of the notebook, the name of the file must be the same as the file provided.
- Upload the notebook only to `http://eol.unibo.it` in the activity specified by the teacher, any other way of submitting the notebook will be ignored

Cooperative work will be heavily sanctioned
The candidate can freely access any kind of materials.